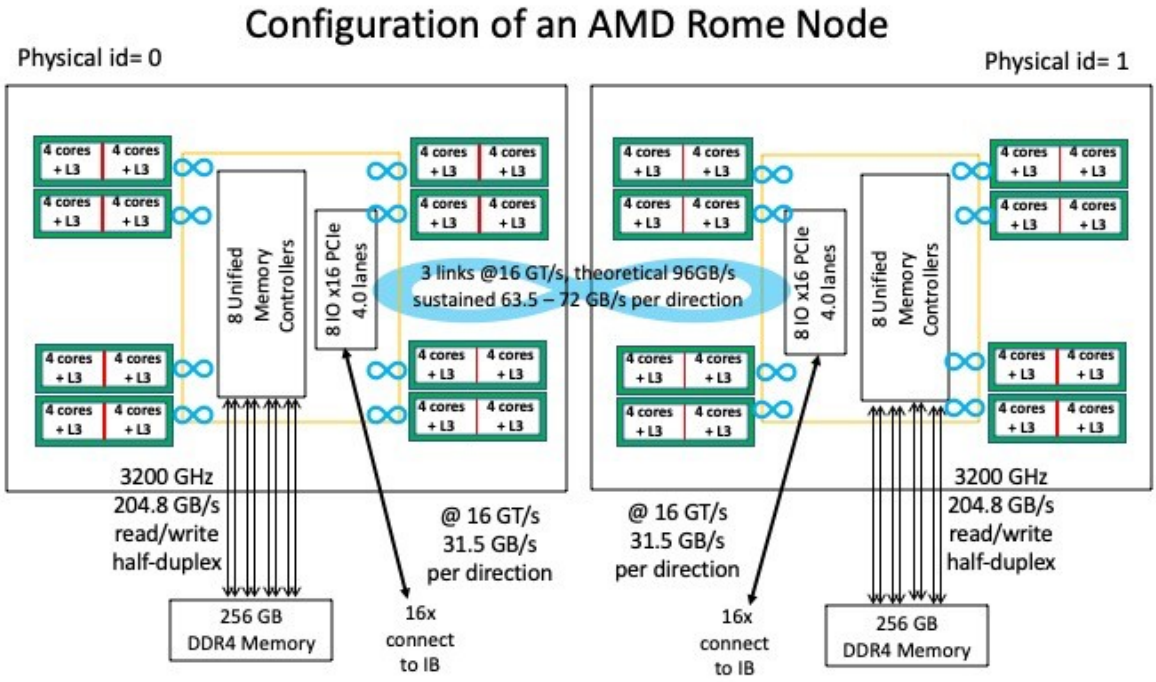
AMD Rome Processors



This is a simplified configuration of an EPYC Rome node with two sockets. Each socket contains eight Core Complex Dies (CCDs, each enclosed in a green box) and one I/O die (IOD, enclosed in a yellow box). The infinity sign (â ¾) represents the Infinity Fabric. Each CCD contains two Core Complexes (CCXs). Each CCX has 4 cores and 16 MB of L3 cache. Thus, there are 64 cores per socket and 128 cores per node.

The AMD EPYC (pronounced "epic") 7742 Rome processor, incorporated into Aitken with the HPE Apollo 9000 system architecture, is in AMD's second generation System-on-Chip (SoC) processor family. Rome has the following high-level features.

- Zen 2 microarchitecture:

  The EPYC 7742 Rome processor has a base CPU clock of 2.25 GHz and a maximum boost clock of 3.4 GHz. There are eight processor dies (CCDs) with a total of 64 cores per socket.
- Hybrid multi-die design:

  Within each socket, the eight processor dies are fabricated on a 7 nanometer (nm) process, while the I/O die is fabricated on a 14 nm process. This design decision was made because the processor dies need the leading edge (and more expensive) 7 nm technology in order to reduce the amount of power and space needed to double the number of cores, and to add more cache, compared to the first-generation EPYC processors. The I/O die retains the less expensive, older 14 nm technology.
- 2nd-generation Infinity Fabric technology:

  Infinity Fabric technology is used for communication among different components throughout the node: within cores, between cores, between CCXs in a CCD, among CCDs in a socket, to the main memory and PCIe, and between the two sockets. The Rome processors are the first x86 systems to support 4th-generation PCIe, which delivers twice the I/O performance (to InfiniBand, storage, NVMe SSD, etc.) over 3rd-generation PCIe.

Processor Hierarchy

The Rome processor hierarchy is as follows:

- Core: A CPU core has private L1I, L1D, and L2 caches, which are shared by two hyperthreads on the core.
- CCX: A core complex includes four cores and a common L3 cache of 16 MB. Different CCXs do not share L3.
- CCD: A core complex die includes two CCXs and an Infinity Link to the I/O die (IOD). The CCDs connect to memory, I/O, and each other through the IOD.
- Socket: A socket includes eight CCDs (total of 64 cores), a common centralized I/O die (includes eight unified memory controllers and eight IO x16 PCIe 4.0 lanesâ total of 128 lanes), and a link to the network interface controller (NIC).
- Node: A node includes two sockets and a network interface controller (NIC). In an Aitken Rome node, the NIC is a dual single-port InfiniBand (IB) High Data Rate (HDR) 200 Gbit/s mezzanine card.

CPU Core

Rome is a 64-bit x86 server microprocessor. A partial list of instructions and features supported in Rome includes **`SSE, SSE2, SSE3, SSSE3, SSE4a, SSE4.1, SSE4.2, AES, FMA, AVX, AVX2`** (256 bit), Integrated x87 FPU (**`FPU`**), Multi-Precision Add-Carry (**`ADX`**), 16-bit Floating Point Conversion (**`F16C`**), and No-eXecute (**`NX`**). For a complete list, run **`cat /proc/cpuinfo`** on a Rome node.

Unlike the Intel Skylake and Cascade Lake processors, Rome does not come with the **`AVX-512`** instructions.

Each Rome core:

- Can sustain execution of four x86 instructions per cycle, using features such as the micro-op cache, advanced branch prediction, and prefetching. The prefetcher works on streaming data and on variable strides, allowing it to accelerate many different data structures.
- Has two 256-bit Fused Multiply-Add (FMA) units and can deliver up to 16 double-precision floating point operations (flops) per cycle. Thus, the peak double-precision flops per node is: 128 cores x 2.25 GHz x 16 = 4.6 TF. With 1,024 Rome nodes in Aitken, the peak is 4.72 PF.
- Can support Simultaneous Multi-threading (SMT), allowing two threads to execute simultaneously per core. SMT can be enabled/disabled in the Basic Input/Output System (BIOS) settings.

  Note: The final setting is yet to be decided. If you plan to take advantage of SMT, use **`cat proc/cpuinfo`** to determine whether it has been enabled.

Linux support for the Zen microarchitecture started with Linux kernel 4.10. The Aitken Rome nodes use Linux kernel 5.3.

The AMD Optimizing C/C++ (AOCC), GNU, Intel, and PGI compilers can be used to compile codes for running on the Rome nodes. To generate optimized x64 code for the Zen 2 microarchitecture, consider using these compiler flags:

- AOCC compiler: **`-march=znver2`**

- Intel compiler: **`-march=core-avx2`** (preferred) or **`-axCORE-AVX2`**
- PGI compiler: **`-tp=zen`**
- GNU compiler (GCC): **`-march=znver2`**

  Note: **`-march=znver2`** is available in the **`gcc 9.x`** compiler and later versions. With **`gcc 8.x`**, use
  `-march=znver1` instead.

For more information, see Compiler Options Quick Reference Guide for AMD EPYC 7xx2 Series Processors.

Cache Hierarchy

The Rome cache hierarchy is as follows:

- op cache (OC): 4K ops, private to each core; 64 sets; 64 bytes/line; 8-way. OC holds instructions that have already been decoded into micro-operations (micro-ops). This is useful when the CPU repeatedly executes a loop of code. Using OC improves:
  - Pipeline latency: because the op cache pipeline is shorter than the traditional fetch and decode pipeline.
  - Bandwidth: because the maximum throughput from the op cache is eight instructions per cycle, whereas the maximum throughput from the traditional fetch and decode pipeline is four instructions per cycle.
  - Power: because there is no need to re-decode instructions.
- L1 instruction cache: 32 KB, private to each core; 64 bytes/line; 8-way. The processor fetches instructions from the instruction cache in 32-byte naturally aligned blocks.
- L1 data cache: 32 KB, private to each core; 64 bytes/line; 8-way; latency: 7-8 cycles for floating point and 4-5 cycles for integer; 2 x 256 bits/cycle load bandwidth to registers; 1 x 256 bits/cycle store bandwidth from registers; write-back policy.

  Note: With the write-back policy, data is updated in the current level cache first. The update in the next level storage is done later when the cache line is ready to be replaced.
- L2 cache: 512 KB, private to each core; unified; inclusive of L1 cache; 64 bytes/line; 8-way; latency: >= 12 cycles; 1 x 256 bits/cycle load bandwidth to L1 cache; 1 x 256 bits/cycle store bandwidth from L1 cache; write-back policy.
- L3 cache: 16 MB shared among four cores in a core complex (CCX); different CCXs do not share L3; total of 256 MB per socket. Within each CCX: 64 bytes/line; 16-way; latency: 39 cycles on average.

  Note: If a core misses in its local L2 and also in the L3, the shadow tags are consulted. If the shadow tag indicated that the data resides in another L2 within the CCX, a cache-to-cache transfer is initiated.
  1 x 256 bits/cycle load bandwidth to L2 of each core; 1 x 256 bits/cycle store bandwidth from L2 of each core; write-back policy; populated by L2 victims.

Memory Subsystem

Unlike the Zen 1 microarchitecture, in which each CCD has its own dual-channel memory controller, the Zen 2 microarchitecture places eight unified memory controllers in the centralized I/O die so that the memory latency is reduced (roughly at 200+ ns) and is more consistent. The memory channels can be split into one, two, or four Non-Uniform Memory Access (NUMA) Nodes per Socket (NPS1, NPS2, and NPS4). The BIOS default for HPE Apollo 9000

systems is NPS4, which is the highest memory bandwidth configuration geared toward HPC applications.

With eight 3,200-GHz memory channels, an 8-byte read or write operation taking place per cycle per channel results in a maximum total memory bandwidth of 204.8 GB/s per socket.

Each memory channel can be connected with up to two Double Data Rate (DDR) fourth-generation Dual In-line Memory Modules (DIMMs). For the Aitken Rome configuration, each channel is connected to a single 32-GB DDR4 registered DIMM (RDIMM) with error correcting code (ECC) support. In total, the amount of memory is 256 GB per socket and 512 GB per node.

Note: Using the one DIMM per Channel (DPC) configuration enables the system to run the memory DIMMs at the highest possible speed; a two-DPC configuration typically requires slightly reduced memory speed.

The memory frequency can be uncoupled, or it can be coupled with the Infinity Fabric frequency through BIOS settings to benefit either bandwidth-bound or latency-bound workloads. The uncoupled mode is used on the Aitken Rome nodes.

- Uncoupled Mode: For throughput-sensitive applications, to obtain higher read/write throughput, the Maximum Memory Bus Frequency option can be set to the maximum allowed (3,200 MT/s). In this case, the Memory Bus will not be synchronized optimally with the slower Infinity Fabric Clock, causing a slight increase in memory access latency.
- Coupled Mode: For latency sensitive applications, memory access latency can be reduced by setting the Maximum Memory Bus Frequency to 2933 MT/s or 2667 MT/s, in order to synchronize with the Infinity Fabric clock.

Intra-Socket Interconnect

The Infinity Fabric, evolved from AMD's previous generation HyperTransport interconnect, is a software-defined, scalable, coherent, and high-performance fabric. It uses sensors embedded in each die to scale control (Scalable Control Fabric, or SCF) and data flow (Scalable Data Fabric, or SDF).

- The SCF uses sensors to monitor die temperature, speed, and voltage across all cores within the dies and controls power management, security, reset, etc.
- The SDF connects the L3 caches to memory and to the configurable I/O lanes. SDF uses the configurable I/O lanes for memory-coherent communications between compute elements on a single die, between different dies on a socket, and between sockets in a node.
- The die-to-die Infinity Fabric bandwidth is 32 bytes for read and 16 bytes for write per Infinity Fabric clock (which has a maximum speed of 1,467 MHz).

Inter-Socket Interconnect

Two EPYC 7742 SoCs are interconnected via Socket to Socket Global Memory Interconnect (xGMI) links, part of the Infinity Fabric that connects all the components of the SoC together. In each Rome node configured with the HPE Apollo 9000 system architecture, there are 3 xGMI links using a total of 48 PCIe lanes. With the xGMI link speed set at 16 GT/s, the theoretical throughput for each direction is 96 GB/s (3 links x 16 GT/s x 2 bytes/transfer) without factoring in the encoding for xGMI, since there is no publication from AMD available. However, the expected efficiencies are 66-75%, so the sustained bandwidth per direction will be 63.5-72 GB/s.

Note: The xGMI link speed and width can be adjusted via BIOS setting. The xGMI Link Max Speed can be set to 10.667, 13, 16 or 18 GT/s. Setting it to a lower speed can save uncore power that can be used to increase core frequency or reduce overall power. It will also decrease cross-socket bandwidth and increase cross-socket latency. xGMI Dynamic Link Width Management saves power during periods of low socket-to-socket data traffic by reducing the number of active xGMI lanes per link from 16 to 8.

Inter-Node Network

There are 1,024 Rome nodes in the Aitken cluster. They are partitioned as follows:

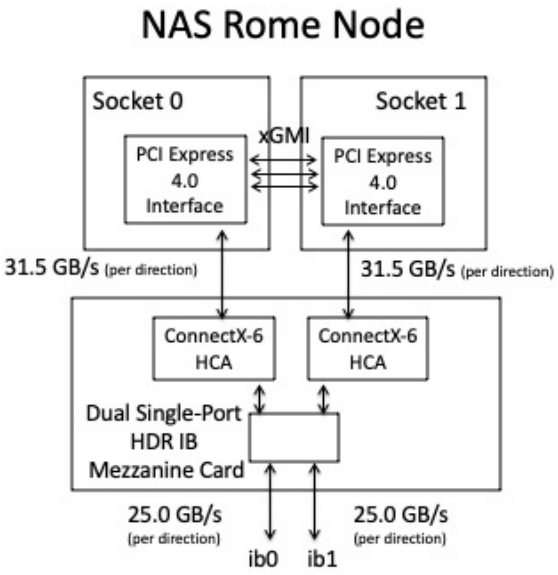
- 8 racks in the cluster (1,024 nodes total)
- 4 enclosures per rack (128 nodes/rack)
- 8 trays (HPE XL925g quad-node tray) per enclosure (32 nodes/enclosure)
- 4 nodes per compute tray

The hostname of each node is r[x1]c[x2]t[x3]n[x4], where x1 = 201 to 208, x2= 1 to 4, x3=1 to 8, and x4 =1 to 4.

The inter-node connection of Aitken's Rome nodes relies on the use of PCIe 4.0 lanes and one dual single-port IB HDR 200 Gbit/s mezzanine card on each node, and two IB HDR 200 Gbit/s premium switches per enclosure. One switch per enclosure facilitates the ib0 fabric, while the other facilitates the ib1 fabric.

As shown in the top diagram of this page, of the 128 PCIe 4.0 lanes per socket, 16 lanes are used to connect a node outwards. With a transfer rate of 16 GT/s, this enables a maximum bandwidth of 31.5 GB/s (when factoring in the 128/130 encoding) for each direction.

The two PCIe 4.0 x16 slots (one from each socket) are connected to the dual single-port IB HDR 200 Gbit/s mezzanine card, as shown below. The mezzanine card contains two separate Mellanox ConnectX-6 Host Channel Adapters (HCA), one for ib0 and the other for ib1, with a bandwidth of 25.0 GB/s per direction and a sub-microsecond latency. Note that the Mellanox ConnectX-6 200-Gbp/s adapters can achieve their maximum bandwidth (25.0 GB/s) in a PCIe 4.0 x16 slot (31.5 GB/s) but not in a PCIe 3.0 x16 slot (15.75 GB/s, as used in NAS Intel Xeon processors).



The 32 nodes in each enclosure are connected to one HPE Apollo 9000 IB HDR 40 Up premium switch (200 Gbit/s) for ib0, and to another for ib1. In each switch, there are two 40-port ConnectX-6 Application-Specific Integrated Circuit (ASIC) chipsâ 80 ports total, of which 32 are used as downlinks to connect to the 32 nodes in the enclosure, 40 are used as uplinks to connect to external targets, and four are used internally to connect between the two ASICs.

For each IB fabricâ with 32 nodes in an enclosure connecting to the same IB switch, forming the first-dimension in the topologyâ the 1,024 Rome nodes form a 6-dimensional enhanced hypercube.

If additional HPE Apollo 9000 systems are added to the cluster, they will form a 7-dimensional enhanced hypercube.

Connection Between Rome and Cascade Lake Nodes

The Rome and Cascade Lake nodes in Aitken will be connected with additional HDR cables on the 8th dimension.

References

- AMD EPYC 7742
- AMD EPYC 7742 specifications
- Compiler Options Quick Ref Guide for AMD EPYC 7xx2 Series Processors (PDF)
- High Performance Computing: Tuning Guide for AMD EPYC 7002 Series Processors (PDF)
- Workload Tuning Guide for AMD EPYC 7002 Series Processor Based Servers (PDF)
- Software Optimization Guide for AMD Family 17h Models 30h and Greater Processors
- AMD Infinity Architecture

Article ID: 658
Last updated: 15 Jun, 2021
Revision: 44
Systems Reference -> Aitken -> AMD Rome Processors
https://www.nas.nasa.gov/hecc/support/kb/entry/658/